PUB-NO: DE019955717A1

DOCUMENT-IDENTIFIER: DE 19955717 A1

TITLE: **Converting unstructured data into structured data** involves suggesting data structure element for selected input data segment that can be structured, allocating structure element as target element

PUBN-DATE: August 24, 2000

INVENTOR-INFORMATION:

NAME	COUNTRY
LEYMANN, FRANK	DE
ROLLER, DIETER	DE

ASSIGNEE-INFORMATION:

NAME	COUNTRY
IBM	US

APPL-NO: DE19955717

APPL-DATE: November 4, 1999

PRIORITY-DATA: EP98121449A (November 11, 1998)

INT-CL (IPC): G06F017/30, G06F017/60

EUR-CL (EPC): G06F017/30

ABSTRACT:

The method is implemented by a computer system and involves a data selection step, in which at least one data segment is selected, whereby this data segment contains part of the input data and can be converted into a structured data segment. At least one data structure element is suggested in a suggestion step. An allocation step involves allocating a data structure element as the target structure element for storing the selected data segment. The segment is then extracted from the input data and stored in the target data structure element. Independent claims are also included for a system for implementing the method, for a data processing program and for a computer program.

(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENT- UND MARKENAMT

(12) **Offenlegungsschrift**

(10) **DE 199 55 717 A 1**

(51) Int. Cl.[7]:
G 06 F 17/30
G 06 F 17/60

(21) Aktenzeichen: 199 55 717.9
(22) Anmeldetag: 4. 11. 1999
(43) Offenlegungstag: 24. 8. 2000

DE 199 55 717 A 1

(30) Unionspriorität:
98121449.7 11. 11. 1998 EP

(71) Anmelder:
International Business Machines Corp., Armonk, N.Y., US

(74) Vertreter:
Duscher, R., Dipl.-Phys. Dr.rer.nat., Pat.-Ass., 71034 Böblingen

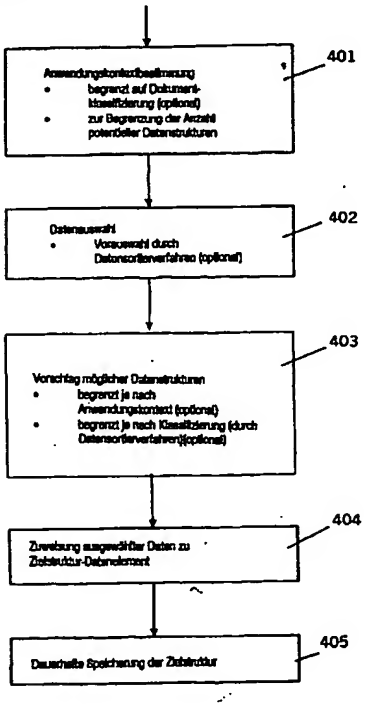
(72) Erfinder:
Leymann, Frank, Dipl.-Math. Dr., 71134 Aidlingen, DE; Roller, Dieter, Dipl.-Phys., 71101 Schönaich, DE

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

Prüfungsantrag gem. § 44 PatG ist gestellt

(54) Umwandlung unstrukturierter Daten in strukturierte Daten

(57) Es wird ein Verfahren zur Umwandlung unstrukturierter Eingabedaten in strukturierte Ausgabedaten beschrieben. Das Verfahren enthält einen Datenauswahlschritt, in dem mindestens ein Datensegment ausgewählt wird, wobei das genannte Datensegment einen Teil der genannten Eingabedaten umfaßt und das genannte Datensegment in ein Datenstrukturelement umgewandelt werden kann. Weiterhin umfaßt das Verfahren der vorliegenden Erfindung einen Vorschlagsschritt, bei dem mindestens ein Datenstrukturelement vorgeschlagen wird. Schließlich umfaßt das Verfahren einen Zuweisungsschritt, bei dem ein Datenstrukturelement als Zieldaten-Strukturelement zur Speicherung des genannten ausgewählten Datensegments zugewiesen wird und bei dem das genannte ausgewählte Datensegment aus den Eingabedaten extrahiert und im genannten Zieldaten-Strukturelement gespeichert wird.



DE 199 55 717 A 1

Beschreibung

1. Hintergrund der Erfindung

1.1 Anwendungsbereich der vorliegenden Erfindung

Die vorliegende Erfindung bezieht sich auf ein Verfahren auf dem Gebiet der Informations-Mining. Genauer gesagt bezieht sich die vorliegende Erfindung auf ein Verfahren zur Behandlung unstrukturierter Eingabedaten.

1.2 Beschreibung und Nachteile der bisherigen Situation

Organisationen erzeugen und erfassen große Datenmengen, die sie in ihren täglichen Abläufen verwenden. Dennoch sind zahlreiche Unternehmen nicht in der Lage, das volle Potential dieser Daten auszuschöpfen, da der Informationsgehalt dieser Daten nicht einfach zu erkennen ist. Die in Verwendung befindlichen Systeme zeichnen Transaktionen genauso auf, wie sie eingehen, also Tag und Nacht, und speichern die Transaktionsdaten in Dateien und Datenbanken ab. Dokumente werden erstellt und in gemeinsamen Dateien oder in von Dokumentverwaltungen bereitgestellten Ablagesystemen abgelegt. Die zunehmende Verbreitung des Internet und seine wachsende weltweite Akzeptanz als Hauptkanal sowohl für die Kommunikation zwischen einzelnen Personen als auch für die Abwicklung von Geschäftsabläufen (beispielsweise durch email) haben die Informationsquellen und somit die Chancen zur Erlangung von Wettbewerbsvorteilen vervielfacht. Business Intelligence Solutions ist ein Begriff, der die Prozesse beschreibt, die insgesamt verwendet werden, um eine bessere Entscheidungsfindung zu erreichen. Die Informations-Mining bezeichnet den Prozeß des Daten-Mining und/oder des Text-Mining. Dabei wird eine moderne Technologie verwendet, mit der wertvolle Einblicke in diese Quellen erreicht werden, die es dem geschäftlichen Benutzer ermöglichen, die richtigen Entscheidungen zu treffen und somit einen Wettbewerbsvorteil zu erlangen, der nötig ist, um in der modernen Wettbewerbsumgebung erfolgreich zu sein. Das Informations-Mining erzeugt aus jeder Quelle im allgemeinen zuvor unbekannte, gut verständliche und belangbare Daten wie beispielsweise Transaktionen, Dokumente, email, Web-Seiten usw. Diese Daten können die Grundlage für wichtige Geschäftsentscheidungen darstellen.

Daten bilden dabei den Rohstoff. Es kann sich hierbei um eine Gruppe diskreter Fakten über Ereignisse handeln, und in diesem Fall spricht man nützlicherweise von strukturierten Aufzeichnungen von Transaktionen, die normalerweise in alphanumerischer Form vorliegen. Doch Dokumente und Web-Seiten sind auch eine Quelle unstrukturierter Daten, die als Bitstrom bereitgestellt und zu Textwörtern und -sätzen einer bestimmten Landessprache dekodiert werden.

Industrieanalysen gehen davon aus, daß unstrukturierte Daten 80% aller Daten in einem Unternehmen ausmachen und nur 20% strukturiert sind; diese Daten haben unterschiedliche Quellen, beispielsweise Text, Bild, Video und Audio. Der überwiegende Anteil der strukturierten Daten liegt allerdings in Textform vor.

Das Daten-Mining nutzt die Infrastruktur gespeicherter Daten (also die Metainformationen der zu verarbeitenden Daten, beispielsweise das Layout der Daten, bestimmte Kennzeichnungen, Beziehungen usw.), um weitere nützliche Informationen zu erlangen. Durch Daten-Mining einer Kundendatenbank könnte man beispielsweise die Erkenntnis gewinnen, daß jeder, der das Produkt A kauft, auch die Produkte B und C kauft, lediglich sechs Monate später.

Nur wenn die von einer Anwendung zu verarbeitenden Eingabedaten eine vordefinierte Struktur einhalten, die in der Anwendung bekannt ist, kann diese Anwendung die Eingabedaten be- und verarbeiten.

Da die Verfügbarkeit strukturierter Daten Voraussetzung für jede weitere Verarbeitung der möglichen Komponenten, die die Bestandteile der unstrukturierten Daten ausmachen, sind, wurde beispielsweise das Text-Mining entwickelt. Text-Mining ist die Anwendung des Prinzips des Daten-Mining auf unstrukturierte oder geringfügig strukturierte Textdateien. Das Text-Mining muß im Gegensatz zum Daten-Mining in einer weniger strukturierten Umgebung erfolgen. Die Dokumente haben nur selten eine starke interne Infrastruktur (und wenn das der Fall ist, dann bezieht sich diese Infrastruktur meistens auf das Dokumentformat und weniger auf den Inhalt). Bei dem Text-Mining werden Metadaten über Dokumente aus den Dokumenten extrahiert. Die Metadaten stellen eine Möglichkeit dar, den Inhalt eines Dokuments anzureichern, und zwar so, daß die Mining-Software dieses Dokument anschließend manipulieren kann. Die Text-Mining Technik ist eine Methode zur Ausweitung des Daten-Mining auf die immensen und immer weiter wachsenden Mengen gespeicherter Texte in einem automatischen Prozeß, in dem strukturierte Daten erstellt werden, die Dokumente beschreiben. Innerhalb des Text-Mining gibt es viele verschiedene Technologien zur Erzeugung von Metadaten für ein Dokument, mit dem Ziel, die Art eines Dokuments zu bestimmen, seine Struktur abzuleiten, usw. Hier einige Beispiele:

Merkmalsextraktion (feature extraction): dient zum Suchen und Extrahieren von Informationen oder Wissen aus Textdokumenten.

Cluster-Technologie (clustering technology): dient zum Sortieren von Dokumenten nach Themen, ermöglicht die Suche nach Schwerpunktthemen in einer Dokumentensammlung usw.

Sämtliche dieser Technologien sind bis zu einem gewissen Grad effektiv und ermöglichen eine Orientierung unter dieser riesigen Anzahl unstrukturierter Informationsquellen. Letztendlich können sie jedoch nicht auf zuverlässige Weise und automatisch den strukturierten Informationsgehalt aus einem unstrukturierten Eingabedokument herausextrahieren. Sie können nur bestimmte Angaben zur Art der Eingabedaten liefern und bieten keine Instrumente zur Umwandlung der unstrukturierten Eingabedaten in strukturierte Eingabedaten an.

1.3 Ziel der vorliegenden Erfindung

Das Prinzip der vorliegenden Erfindung beruht auf dem Ziel, ein Verfahren zum Herausfiltern strukturierter Daten aus einer unstrukturierten Eingabe bereitzustellen und auf diese Weise eine Anwendung zu unterstützen, die für ihre Verarbeitung strukturierten Eingabedaten benötigt.

2. Zusammenfassung und Vorteile der vorliegenden Erfindung

Die Ziele der vorliegenden Erfindung werden gemäß der Ausführung von Anspruch 1 erreicht.

Das Prinzip der vorliegenden Erfindung bezieht sich auf ein von einem Computersystem ausgeführten Verfahren zur Umwandlung unstrukturierter Eingabedaten in strukturierte Ausgabedaten. Das Verfahren der vorliegenden Erfindung umfaßt einen Schritt der Datenauswahl, bei dem mindestens ein Datensegment ausgewählt wird, wobei das genannte Datensegment einen Teil der genannten Eingabedaten umfaßt und das genannte Datensegment in ein Datenstrukturelement umgewandelt werden kann. Das Verfahren der vorlie-

genden Erfindung umfaßt weiterhin einen Schritt, bei dem mindestens ein Datenstrukturelement vorgeschlagen wird. Schließlich umfaßt das Verfahren der vorliegenden Erfindung einen Schritt der Zuweisung, bei dem ein Datenstrukturelement als Zieldatenstrukturelement zur Speicherung des genannten ausgewählten Datensegments zugewiesen wird und bei dem das genannte ausgewählte Datensegment aus den genannten Eingabedaten extrahiert und im genannten Zieldatenstrukturelement gespeichert wird.

Das Prinzip der vorliegenden Erfindung ermöglicht es, die riesigen immer weiter wachsenden Mengen unstrukturierter elektronischer Daten zu bewältigen. Aus Sicht des Benutzers besteht der Vorteil der vorliegenden Erfindung darin, die Aufgabe der Extraktion von Daten aus unstrukturierten Eingabedaten für Anwendungen zu vereinfachen, die strukturierte Daten erwarten. Die für die Datenextraktion benötigte Zeit wird deutlich reduziert, und das fehleranfällige Abtippen wird nicht länger benötigt. Der Benutzer kann ein Datensegment für die Extraktion beliebig auswählen und ist dabei nicht durch das System eingeschränkt. Der Benutzer muß nicht länger im voraus die potentiellen Datenstrukturen kennen. Stattdessen bietet das Verfahren der vorliegenden Erfindung im Vorschlagsschritt die möglichen Datenstrukturen zur Auswahl an.

Entsprechend einem weiteren Ausführungsbeispiel der vorliegenden Erfindung enthält das Verfahren auch einen Schritt zur Speicherung, in dem das genannte Zieldatenstrukturelement dauerhaft gespeichert wird.

Die dauerhafte Speicherung der erfaßten Eingabedaten ermöglicht es, daß jede beliebige Anwendung zu einem späteren Zeitpunkt darauf zugreifen kann.

Entsprechend einem weiteren Ausführungsbeispiel der vorliegenden Erfindung werden im Vorschlagsschritt Datenstrukturelemente und/oder Datenstrukturen vorgeschlagen, wobei die genannten Datenstrukturen ein oder mehrere Datenstrukturelemente und/oder ein oder mehrere weitere Datenstrukturen enthalten.

Dieses weitere Ausführungsbeispiel der vorliegenden Erfindung vermeidet Beschränkungen hinsichtlich dem Layouts der beteiligten Datenstrukturen. Jede Datenstruktur kann sich aus atomischen Datenelementen und/oder zusätzlichen Datenstrukturen (mit derselben Substruktur) zusammensetzen. Das Prinzip der vorliegenden Erfindung erlegt hinsichtlich des rekursiven Layouts keine Beschränkungen auf.

Entsprechend einem weiteren Ausführungsbeispiel der vorliegenden Erfindung geht dem Schritt der Datenauswahl ein Schritt zur Bestimmung des Anwendungskontextes voraus, bei dem mindestens eine Zielanwendung festgelegt wird, um gegebenenfalls die strukturierten Ausgabedaten zu verarbeiten. Im Schritt zur Bestimmung des Anwendungskontextes können die genannten Eingabedaten automatisch vom genannten Computersystem klassifiziert und mindestens einer Zielanwendung zugewiesen werden. Ersatzweise oder zusätzlich kann im genannten Schritt zur Bestimmung des Anwendungskontextes ein Benutzer aus einer Gruppe von Anwendungen mindestens eine Zielanwendung auswählen. Schließlich werden im genannten Vorschlagsschritt nur solche Datenstrukturelemente vorgeschlagen, die sich auf die genannte Zielanwendung beziehen.

Die Möglichkeit, einen Anwendungskontext auszuwählen, gestattet eine deutliche Reduzierung potentieller Zieldatenstrukturen im Vorschlagsschritt. Die Klassifizierung der Eingabedaten auf automatische Weise führt zu weiteren Vorteilen. Die Klassifizierung kann in einem automatisch zugewiesenen oder einem zuvor ausgewählten Anwendungskontext resultieren. Im letzteren Fall läßt sich der zuvor ausgewählte Anwendungskontext vom Benutzer weiter verfeinern.

Entsprechend einem weiteren Ausführungsbeispiel der vorliegenden Erfindung analysiert im genannten Datenauswahlschritt ein Parser die genannten Eingabedaten, klassifiziert potentielle Datensegmente und wählt im voraus Datensegmente aus, die für eine Zielanwendung möglicherweise relevant sind.

Auf der Basis dieser Vorgehensweise vereinfacht das Prinzip der vorliegenden Erfindung auch den Auswahlprozeß. Das Verfahren schlägt im voraus ausgewählte Datensegmente vor, die ein Benutzer übernehmen oder aufgrund seines zusätzlichen Wissens ergänzen kann.

Entsprechend einem weiteren Ausführungsbeispiel der vorliegenden Erfindung wird vorgeschlagen, das Verfahren in einer Zielanwendung und/oder in einem Mailing-System und/oder in einem Textverarbeitungsprogramm zu integrieren.

Der Vorteil besteht darin, daß das Verfahren an denjenigen Stellen in einem System verfügbar gemacht wird, wo unstrukturierte Daten im System eingehen. Deshalb findet die Umwandlung in strukturierte Daten so früh wie möglich statt, was es allen Anwendungen, die zu einem späteren Zeitpunkt ausgeführt werden, ermöglicht, von den strukturierten Daten zu profitieren.

Kurze Beschreibung der Zeichnungen

Fig. 1 zeigt ein Beispiel einer manuellen Datenerfassung unter Verwendung von Formularen gemäß dem Stand der Technik.

Fig. 2 ist eine Abbildung der Erfassung strukturierter Ausgabedaten aus unstrukturierten Eingabedaten in Übereinstimmung mit der vorliegenden Erfindung.

Fig. 3 veranschaulicht das weitere Ausführungsbeispiel eines Anwendungskontextes zur Begrenzung der Gruppe potentieller Zieldatenstrukturen.

Fig. 4 ist eine Zusammenfassung des Verfahrens der vorliegenden Erfindung.

4. Beschreibung des bevorzugten Ausführungsbeispiels

Wenn in der Beschreibung der vorliegenden Erfindung von elektronischen Daten oder einem elektronischen Dokument usw. die Rede ist, dann sind damit alle Datenarten gemeint.

4.1 Einführung

Die im Einsatz befindlichen Systeme (Systeme also, die die täglichen Abläufe eines Unternehmens steuern) arbeiten mit strukturierten Daten. Die Zusammensetzung solcher Daten aus einfachen atomischen Datentypen (in einfachen Fällen aus Ganzzahlen, Strings usw.) ist vordefiniert und in den Systemen, die diese Daten verarbeiten, bekannt. Ohne solche Metadaten funktioniert keine der klassischen Anwendungen oder gar Algorithmen: Daten, die nicht strukturiert wurden, können grundsätzlich nicht verarbeitet werden (zumindest nicht hinsichtlich ihrer potentiellen Bestandteile).

Wenn Menschen miteinander kommunizieren, werden Daten hauptsächlich in 'unstrukturierter' Form verwendet.

Beispiele für solche Daten sind Text, Bild und Sprache, die zwischen Menschen ausgetauscht werden, die in Briefen, Telefax-Nachrichten, e-mails, Telefongesprächen usw. miteinander kommunizieren. Diese Daten besitzen keine Struktur, die für die Anwendungen oder Algorithmen verfügbar ist. Folglich muß der Mensch aus diesen unstrukturierten Eingaben die relevanten Daten herausfiltern und sie entsprechend den Anforderungen der Anwendung struktu-

rieren, wenn die unstrukturierte Eingabe sonst auf die Abläufe in einem Unternehmen negative Auswirkungen hätte.

Fig. 1 zeigt, was heutzutage in solchen Situationen normalerweise gemacht wird: Die unstrukturierte Eingabe (100) (in diesem Fall ein Textverarbeitungsanhang an einem e-mail) wird von einem Menschen gelesen. Der Mensch versteht den Brief und weiß, welche Art von Daten die Zielanwendung benötigt. Der Mensch filtert also die erforderlichen Daten aus der unstrukturierten Eingabe heraus und gibt sie (Feld für Feld) in ein Formular (101) ein – Schritt 1 in Fig. 1. Dieses Formular könnte beispielsweise bereits in der Schnittstelle der Zielanwendung dargestellt werden. Sobald das Formular vollständig ausgefüllt ist, teilt er dies der Anwendung mit, die dann die inzwischen strukturierte Eingabe dazu verwendet, beispielsweise eine Datenbank (oder eine Datei usw.) (102) zu manipulieren – Schritt 2 in Fig. 1.

Diese Vorgehensweise ist nicht nur mühsam und zeitaufwendig, sondern bekanntermaßen auch fehleranfällig: Der Mensch muß sich an die Daten aus der unstrukturierten Eingabe erinnern (beispielsweise an den Namen eines Kunden und seine Schreibweise) und sie in das Formular der Anwendung eingeben. Natürlich kann er diese Daten auch handschriftlich auf ein Blatt Papier schreiben oder einfach nur die Datenquelle und das Zielformular in zwei verschiedenen Fenstern gleichzeitig auf dem Bildschirm anzeigen lassen, doch bleiben die Fehleranfälligkeit und der Aufwand hoch.

4.2 Die Lösung

Die Lösung in Übereinstimmung mit der vorliegenden Erfindung wird in Fig. 2 dargestellt.

Das Prinzip der vorliegenden Erfindung beruht auf der Verwendung von Menüs, die die Erfassung strukturierter Daten anhand von unstrukturierten Eingaben unterstützen. Solche Menüs könnten beispielsweise in Form von stufenweisen Kontextmenüs angeboten werden. Zu diesem Zweck könnte die Software, die zum Durchforsten der unstrukturierten Eingabe (201) verwendet wird (beispielsweise ein Textverarbeitungsprogramm), durch eine Implementierung des vorgeschlagenen Verfahrens in Übereinstimmung mit der vorliegenden Erfindung erweitert werden.

Das beschriebene Verfahren gestattet in einem Auswahlschritt die Auswahl (beispielsweise durch Markierung) eines Teils der Eingabe (202), die von einem Menschen als relevant für eine Anwendung angegeben wurde, die eine strukturierte Eingabe erfordert. Als zusätzliche Ergänzung könnte das Verfahren einzelne Elemente des Daten-Minings verwenden. Wenn man beispielsweise die "Merkmalsextraktionstechnologie" und die "Klassifikationstechnologie" einsetzt, kann ein Parser potentielle Datensegmente, die für eine Zielanwendung relevant sind, automatisch erkennen und klassifizieren. Auf der Grundlage dieses Parser-Schritts könnten erkannte Datensegmente, die möglicherweise relevant sind, bereits durch das beschriebene Verfahren im voraus ausgewählt werden. Ein Benutzer könnte daraufhin während des Verfahrensablaufs diese im voraus ausgewählten Datensegmente verwenden oder Ergänzungen in die durch dieses Verfahren beschriebene Vorauswahl einfügen.

Als nächstes wird in einem Vorschlagsschritt ein Kontextmenü geöffnet, das alle möglichen für die Datenarten relevanten Datenstrukturen der Zielanwendung auflistet (Gesamtmenü (203) in Fig. 2, Schritt 1); die Öffnung des Kontextmenüs könnte in der heute für Textverarbeitungsprogramme üblichen Weise erfolgen, indem man beispielsweise die rechte Maustaste drückt, während man mit dem Mauspfeil auf den markierten Menüpunkt zeigt. Es sei darauf hingewiesen, daß es hierzu verschiedene Alternativen gibt (die sich auch kombinieren lassen):

Das Gesamtmenü könnte alle Datenstrukturen aufführen, die für das Unternehmen relevant sind.

Das Gesamtmenü könnte alle Datenstrukturen aufführen, die für eine bestimmte Zielanwendung oder eine bestimmte Gruppe von Zielanwendungen relevant sind.

Wenn man das Klassifizierungsergebnis eines Parsers nutzt, der die Eingabedaten analysiert hat, könnte das Gesamtmenü alle Datenstrukturen aufführen, die für den Dokumenttyp relevant sind, zu dem die Eingabedaten gehören.

An dieser Stelle wird die Zieleingabestruktur ausgewählt und die Auflistung der Datenstrukturen der nächsten Ebene (204), die das Ziel darstellen, angezeigt (Schritt 2 in Fig. 2). Der Einfachheit halber gehen wir davon aus, daß die letzteren Datenstrukturen bereits atomisch sind (sich selbst also nicht in weitere Subelemente abbauen lassen), so daß keine weitere Verfeinerung notwendig ist: Deshalb wird der zweite Kasten in Fig. 2 als Attributmenü bezeichnet. Ansonsten geht die Verfeinerung durch Öffnen zusätzlicher Listenkästen weiter, das heißt, die Strukturelemente, die selbst eine Struktur bilden, könnten eine Struktur mit weiteren Strukturelementen darstellen, usw. Durch Auswahl eines Menüpunkts aus dem Attributmenü wird der ausgewählte Teil aus der unstrukturierten Eingabe als Wert für dieses Attribut zugeordnet, was den Zuweisungsschritt des aktuellen Verfahrens vervollständigt.

Natürlich wäre es auch möglich, die potentiellen Datenstrukturen und die Teilsubstrukturen in einem einzigen Dialog darzustellen. Dies wirkt sich lediglich auf die Verwendbarkeit des beschriebenen Verfahrens aus. Die Entscheidung, ob man eine Abstufung verwendet, hängt von der Komplexität der beteiligten Datenstrukturen ab.

Auf diese Weise werden Datenstrukturen, die den Eingabeformularen der Zielanwendungen entsprechen, gefüllt (das heißt, als Instanz behandelt). Die Instanzen dieser strukturierten Daten könnten im Speicher aufbewahrt werden (Flüchtiger Cache (205), Schritt 3 in Fig. 2), bis der Mensch angibt, daß alle erforderlichen Daten erfaßt wurden.

Als nächstes werden die Cache-Instanzen an die Zielanwendung weitergeleitet (Schritt 4 in Fig. 2), um die erfaßten und strukturierten Ausgabedaten dauerhaft zu speichern (206). Durch diesen Speicherschritt wird das Verfahren vervollständigt.

Um die Verwendbarkeit der Datenerfassung weiter zu verbessern, schlägt die Beschreibung der vorliegenden Erfindung vor, die im Gesamtmenü (203) angezeigte Liste in Untergruppen zu unterteilen, indem der geeignete Anwendungskontext ausgewählt wird. Unter einem Anwendungskontext kann man sich eine oder mehrere Anwendungen vorstellen, die in der Lage sind, Daten zu verarbeiten, die in der unstrukturierten Eingabe enthalten sind. Wie in Fig. 3 dargestellt ist, könnte diese Auswahl dadurch unterstützt werden, daß man in die Menüleiste der Software, die zum Durchsuchen der unstrukturierten Eingabe (301) dient, ein Anwendungskontextmenü hinzufügt.

Bei der Auswahl des Anwendungskontextmenüs wird eine Liste (302) der verfügbaren Gesamtmenüs angezeigt. Wenn ein Menüpunkt aus dieser Liste ausgewählt wird, wird nur die entsprechende Untergruppe an Datenstrukturen im oben beschriebenen Gesamtmenü (203) angezeigt. Dadurch wird der Schritt zur Bestimmung des Anwendungskontextes in Übereinstimmung mit der vorliegenden Erfindung vervollständigt. Ein Nebeneffekt der Auswahl eines Menüpunkts aus dem Anwendungskontextmenü ist, daß die Browser-Software weiß, daß nicht die Standard-Kontextmenüs (also das Textmenü, wenn es sich beim ausgewählten Menüpunkt um einen Textteil in einem Textverarbeitungsprogramm handelt) für ausgewählte Teile der unstrukturier-

ten Eingabe Pop-Up-Menüs sein müssen. Stattdessen werden die Kontextmenüs, die zur ausgewählten Anwendung gehören, angezeigt.

Wie beim Vorschlagsschritt ist auch hier anzumerken, daß hinsichtlich dieser Komponente des Verfahrens verschiedene andere Möglichkeiten existieren (die sich ebenfalls kombinieren lassen):

Das Anwendungskontextmenü könnte alle Anwendungskontexte aufführen, die für das Unternehmen relevant sind.

Wenn man das Klassifizierungsergebnis eines Parsers nutzt, der die Eingabedaten analysiert hat, könnte das Anwendungskontextmenü nur diejenigen Anwendungskontexte aufführen, die für den Dokumenttyp relevant sind, zu dem die Eingabedaten gehören.

Das Prinzip der vorliegenden Erfindung ließe sich so umsetzen, daß die Ersteller von Anwendungen Komponenten (beispielsweise Java Beans) bereitstellen, die Anwendungskontexte oder Kontextmenüs kreieren. Auf der Grundlage dieser Komponenten könnte die Browser-Software dann die Menüleisten und Kontextmenüs zusammensetzen. Wenn die resultierende Browser-Software auch solche Komponenten (beispielsweise über Referenzierung) einschließt, dann unterstützt die Software sofort die Datenerfassung.

Zusammenfassend enthält das beschriebene Verfahren die folgenden Schritte, die in **Fig. 4** dargestellt sind:

1. In einem optionalen Anwendungskontextbestimmungsschritt **(401)** wird eine oder mehrere Anwendungen, die in der Lage wären, die unstrukturierten Daten zu verarbeiten, ausgewählt. Die Anzahl der möglichen und beschriebenen Anwendungskontexte kann durch eine moderne Dokumentklassifizierung vordefiniert sein oder dynamisch bestimmt werden. Im Verlauf des beschriebenen Verfahrens wird dieser ausgewählte Anwendungskontext dazu verwendet, die Anzahl der möglichen Datenstrukturen in den folgenden Schritten zu begrenzen.
2. Im nächsten Schritt, dem Datenauswahlschritt **(402)**, können Datensegmente innerhalb der Eingabe als Elemente einer Datenstruktur ausgewählt werden. Als mögliche Erweiterung kann ein Parser, der eine Merkmalsextraktion besitzt, verwendet werden, um die Eingabe mit bereits ausgewählten Datensegmenten vorzuverarbeiten.
3. Im Vorschlagsschritt **(403)** werden mögliche Datenstrukturen vorgeschlagen, die das ausgewählte Datensegment enthalten könnten. Die Gruppe der vorgeschlagenen Datenstrukturen könnte durch den Anwendungskontext oder das Klassifizierungsergebnis eines Parsers, der Daten-Mining auf der Basis des vollständigen unstrukturierten Eingabedokuments oder des ausgewählten Datensegments einsetzt, eingegrenzt werden.
4. Im Zuweisungsschritt **(404)** wird ein Zieldaten-Strukturelement zugewiesen und zur Speicherung des genannten ausgewählten Datensegments verwendet. Das Verfahren extrahiert das ausgewählte Datensegment aus den genannten Eingabedaten und speichert es im Zieldaten-Strukturelement.
5. Zum Schluß werden die erfaßten und strukturierten Ausgabedaten dauerhaft abgespeichert; das heißt, das Verfahren wird durch den Speicherschritt abgeschlossen **(405)**.

4.3 Vorteil

Immer mehr Kommunikation wird durch Geräte und Software unterstützt, die unstrukturierte Daten (wie beispielsweise Text) erzeugen. Zum Beispiel verbreitet sich e-mail zusehends: Nicht nur Versorgungsketten sondern auch Kundenwertschöpfungsketten und Kundenbetreuungssysteme werden davon berührt. Andererseits erwarten bestehende Anwendungen, die die jeweiligen Geschäftsprozesse bereits unterstützen, strukturierte Daten. Es fehlt also eine Übereinstimmung zwischen der Informationsquelle, die unstrukturierte Daten erzeugt, und dem Informationsziel (Datenverarbeitungsanwendungen), das strukturierte Daten benötigt.

Aus der Sicht eines Benutzers besteht der Vorteil der vorliegenden Erfindung darin, das Extrahieren von Daten aus unstrukturierten Eingaben für solche Anwendungen, die strukturierte Daten erwarten, zu vereinfachen: Die Zeit für die Datenextraktion wird verringert, und Fehler durch erneute Eingabe werden vermieden. Dies setzt sich direkt in Einsparungen für den Arbeitgeber des Benutzers um.

Aus der Sicht des Programmierers einer Software für das Durchsuchen unstrukturierter Daten bietet das Prinzip der vorliegenden Erfindung eine einfache Möglichkeit, ein Verfahren zur Datenerfassung umzusetzen.

Patentansprüche

1. Ein Verfahren, das von einem Computersystem ausgeführt wird, zur Umwandlung unstrukturierter Eingabedaten in strukturierte Ausgabedaten,
wobei das genannte Verfahren einen Datenauswahlschritt umfaßt, in dem mindestens ein Datensegment ausgewählt ist, wobei das genannte Datensegment einen Teil der genannten Eingabedaten enthält und das genannte Datensegment in ein Datenstrukturelement umgewandelt werden kann; und
wobei das genannte Verfahren einen Vorschlagsschritt umfaßt, in dem mindestens ein Datenstrukturelement vorgeschlagen wird; und
wobei das genannte Verfahren einen Zuweisungsschritt umfaßt, in dem ein Datenstrukturelement als Zieldaten-Strukturelement zur Speicherung des genannten ausgewählten Datensegments zugewiesen wird, und wobei das genannte ausgewählte Datensegment aus den genannten Eingabedaten extrahiert und im genannten Zieldaten-Strukturelement gespeichert wird.
2. Verfahren zur Umwandlung unstrukturierter Eingabedaten in strukturierte Ausgabedaten gemäß Anspruch 1, wobei das genannte Verfahren durch einen Speicherschritt abgeschlossen wird, in dem das genannte Zieldaten-Strukturelement dauerhaft abgespeichert wird.
3. Verfahren zur Umwandlung unstrukturierter Eingabedaten in strukturierte Ausgabedaten gemäß Anspruch 1 oder 2, wobei im genannten Vorschlagsschritt entweder Datenstrukturelemente und/oder
Datenstrukturen vorgeschlagen werden, wobei die genannten Datenstrukturen ein oder mehrere Datenstrukturelemente und/oder eine oder mehrere weitere Datenstrukturen umfassen.
4. Verfahren zur Umwandlung unstrukturierter Eingabedaten in strukturierte Ausgabedaten gemäß Anspruch 1, 2 oder 3,
wobei dem genannten Datenauswahlschritt ein Anwendungskontextbestimmungsschritt zur Bestimmung mindestens einer Zielanwendung zur potentiellen Verarbeitung der genannten strukturierten Ausgabedaten vorausgeht,
wobei im genannten Anwendungskontextbestimmungsschritt die genannten Eingabedaten automatisch

vom genannten Computersystem klassifiziert und mindestens einer Zielanwendung zugewiesen werden; und/oder

wobei im genannten Anwendungskontextbestimmungsschritt ein Benutzer aus einer Gruppe von Anwendungen mindestens eine Zielanwendung auswählt; und wobei im genannten Vorschlagsschritt nur solche Datenstrukturelemente vorgeschlagen werden, die mit der genannten Zielanwendung zusammenhängen.

5. Verfahren zur Umwandlung unstrukturierter Eingabedaten in strukturierte Ausgabedaten gemäß Anspruch 1, 2, 3 oder 4,

wobei im genannten Datenauswahlschritt ein Parser die genannten Eingabedaten analysiert, dieser Parser potentielle Datensegmente klassifiziert und Datensegmente, die möglicherweise für eine Zielanwendung relevant sind, im voraus auswählt.

6. Verfahren zur Umwandlung unstrukturierter Importdaten in strukturierte Ausgabedaten gemäß Anspruch 1, 2, 3, 4 oder 5,

wobei das genannte Verfahren in einer Zielanwendung integriert ist; und/oder

wobei das genannte Verfahren in einem Mailing-System integriert ist; und/oder

wobei das genannte Verfahren in einem Textverarbeitungsprogramm integriert ist.

7. Ein System mit einem Mittel, das darauf ausgelegt ist, die Schritte entsprechend dem Prinzip der vorliegenden Erfindung gemäß einem der Ansprüche 1 bis 6 auszuführen.

8. Ein Datenverarbeitungsprogramm, das in einem Datenverarbeitungssystem ausgeführt werden kann und Software-Codeteile zur Ausführung eines Verfahrens gemäß einem der Ansprüche 1 bis 6 enthält.

9. Ein Computerprogramm, das auf einem Speichermedium gespeichert ist, das von einem Computer gelesen werden kann, wobei dieses Programm mit einem computerlesbaren Programmittel kombiniert ist, welches einen Computer veranlaßt, ein Verfahren gemäß einem der Ansprüche 1 bis 6 auszuführen.

Hierzu 4 Seite(n) Zeichnungen

Nummer: DE 199 55 717 A1
Int. Cl.[7]: G 06 F 17/30
Offenlegungstag: 24. August 2000

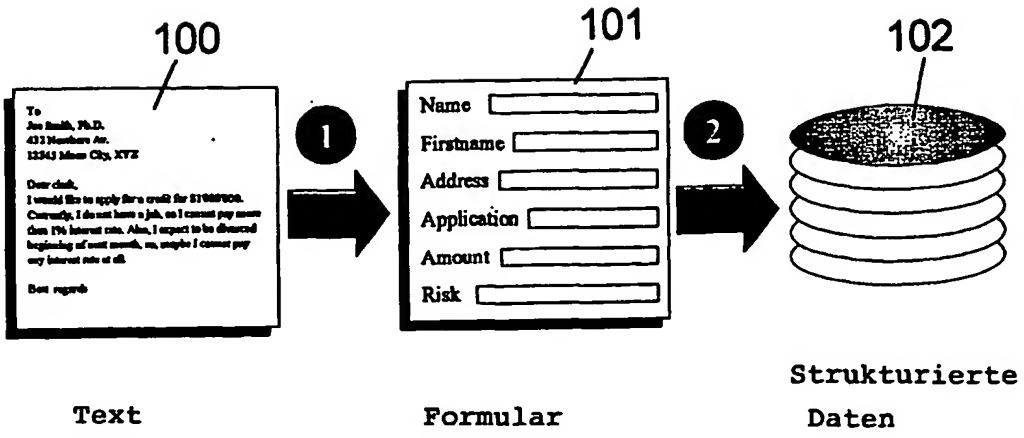


FIG. 1

Nummer: DE 199 55 717 A1
Int. Cl.[7]: G 06 F 17/30
Offenlegungstag: 24. August 2000

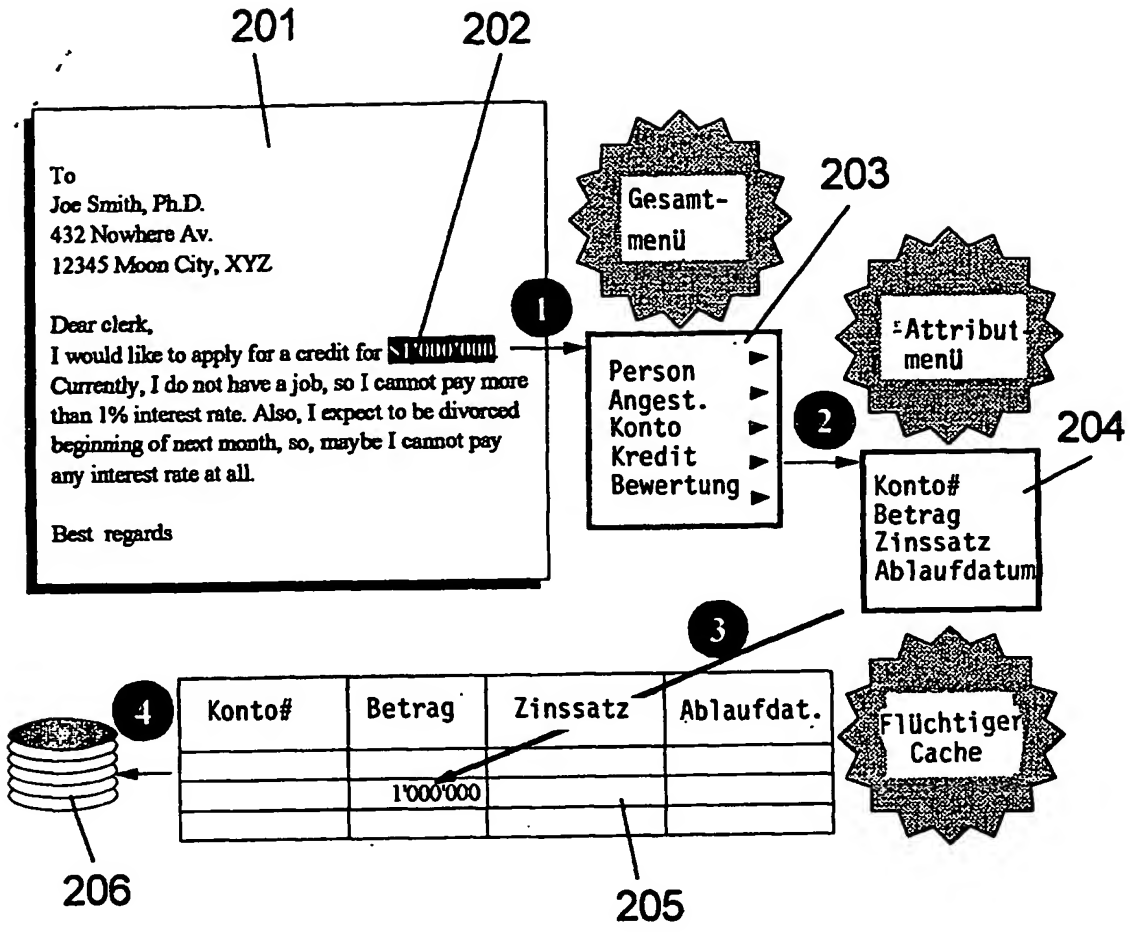


FIG. 2

Nummer: DE 199 55 717 A1
Int. Cl.[7]: G 06 F 17/30
Offenlegungstag: 24. August 2000

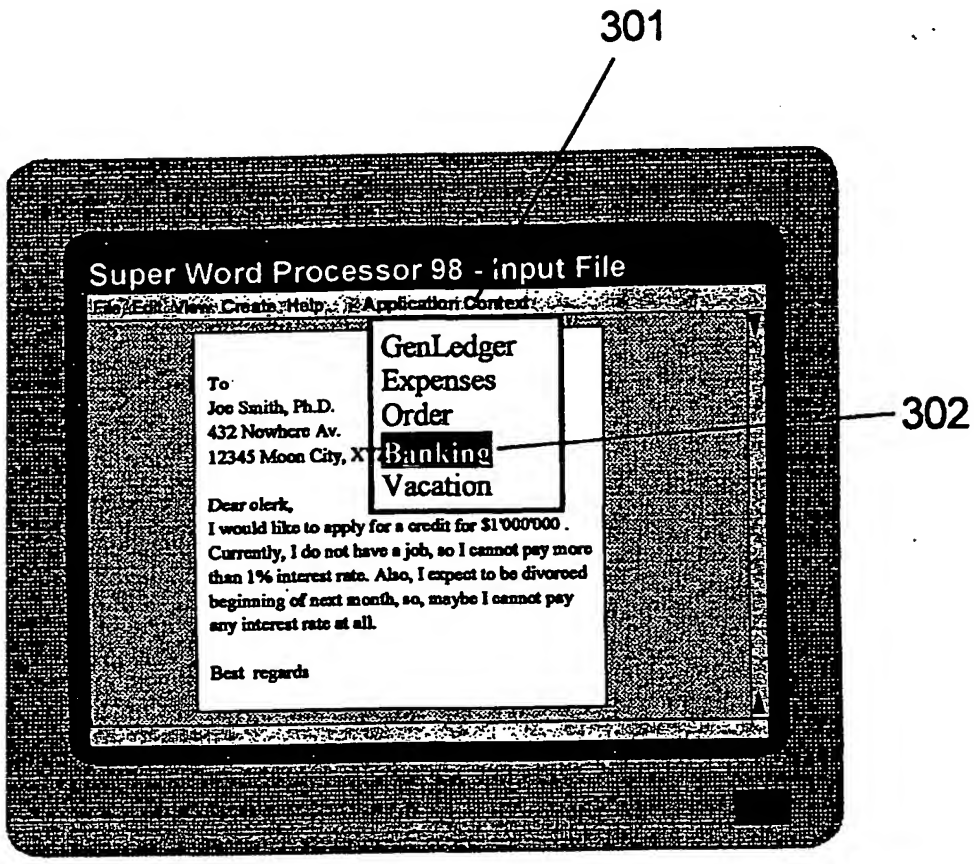


FIG. 3

Nummer: DE 199 55 717 A1
Int. Cl.[7]: G 06 F 17/30
Offenlegungstag: 24. August 2000

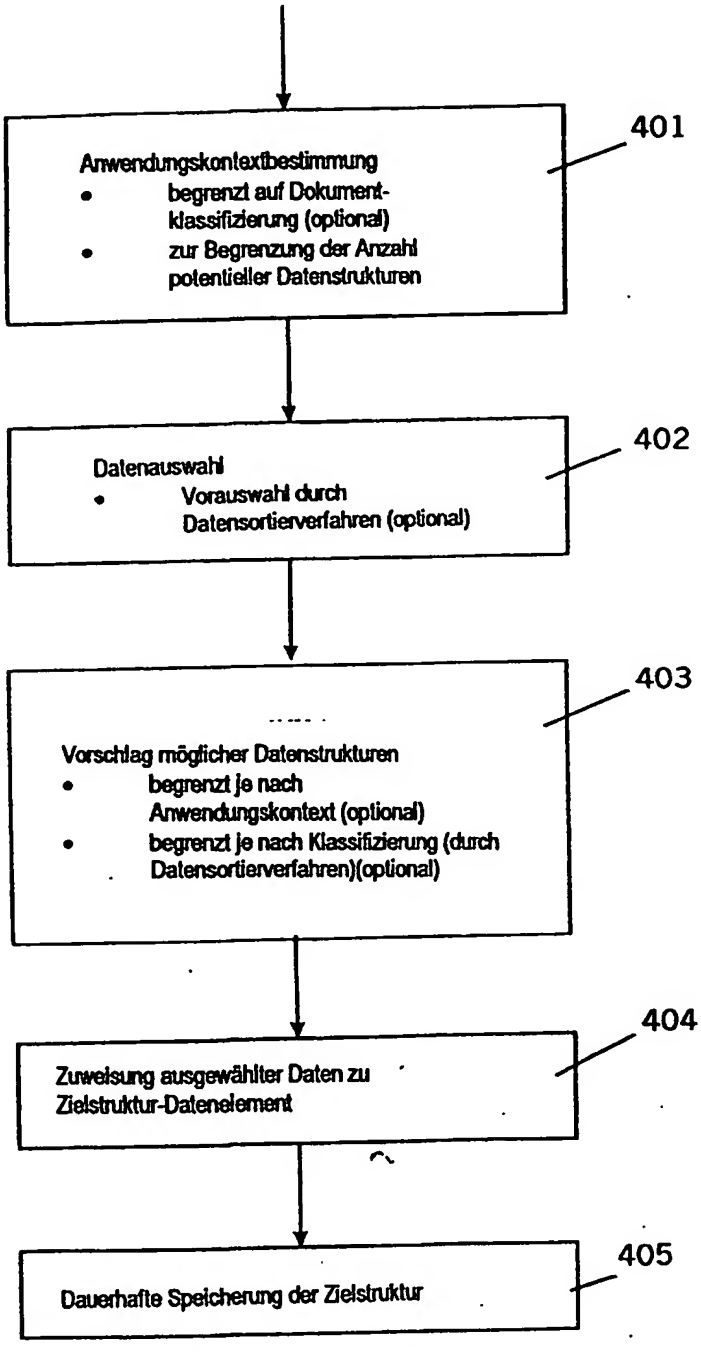


FIG. 4